All together now: Simultaneous Object Detection and Continuous Pose Estimation using a Hough Forest with Probabilistic Locally Enhanced Voting

Carolina Redondo-Cabrera[1]
carolina.redondoc@alu.uah.es
Roberto López-Sastre[1]
robertoj.lopez@uah.es
Tinne Tuytelaars[2]
Tinne.Tuytelaars@esat.kuleuven.be

[1] University of Alcalá
GRAM
Alcalá de Henares, ES

[2] K.U. Leuven,
ESAT-PSI, iMINDS
Leuven, BE

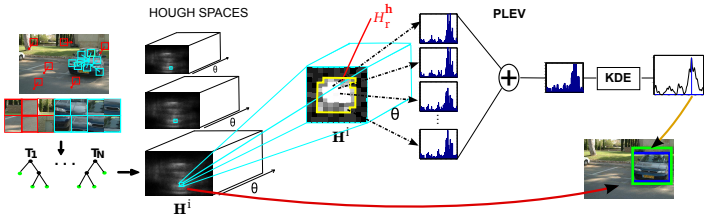


Figure 1: Our approach is able to jointly estimate the localization and the continuous pose of objects. To this end, we follow a HF regression voting in conjunction with our PLEV strategy, to integrate votes from a local region in the Hough space near the detected modes.

Object category detection has received a lot of attention over the last decades. Recently, several approaches have gone one step further proposing solutions for the problem of simultaneous object category detection and pose estimation [1, 3, 5]. In this paper, we tackle this problem using Hough Forests (HF) [2]. We propose a new approach (see Figure 1) which *jointly* solves both tasks, providing detection hypotheses and *probabilistic* estimates of their *continuous* pose.

We first introduce a new formulation for the regression to be performed with HF, incorporating an **uncertainty criterion for the continuous pose of the categories**. This uncertainty in pose is decoupled from the traditional localization uncertainty [2], which allows us to randomly choose between them during the HF learning. The resulting HF can effectively locate objects and estimate their pose.

For a set of patches S, we formulate this pose uncertainty as follows,

$$\mathcal{M}_p(S) = \sum_{child \in (left,right)} \sum_{j:c_j=1} \left(\frac{min\{(||\theta_j - \boldsymbol{\theta}_A||), 360^\circ - (||\theta_j - \boldsymbol{\theta}_A||)\}}{180^\circ} \right)^2, \quad (1)$$

where c_j is the class label of the j patch ($c_j = 1$ for foreground patches, and $c_j = 0$ for background patches), θ_j encodes the continuous pose annotation for the patch j, and $\boldsymbol{\theta_A}$ is the viewpoint angle average over all foreground patches in the set of patches S^{child}. Randomly switching between this pose uncertainty and the localization uncertainty of [2] guarantees that the leaves of our decision trees gather image patches which vote not only for a similar object localization, but also for a similar pose.

However, the extension of the Hough space to cover also the pose regression turns out to be suboptimal. The main reason is that the pose voting is very noisy, as we have experimentally observed, especially for views with shared appearance (*e.g.* think of a frontal vs. frontal-left views of a car). Instead, we propose to first localize the object, and then estimate its pose. For this second step, a novel regression strategy is introduced, named **Probabilistic Locally Enhanced Voting** (PLEV), which consists in modulating the regression with a kernel density estimation (KDE) to consolidate all the votes in a *local* Hough region near the maxima detected in the Hough space.

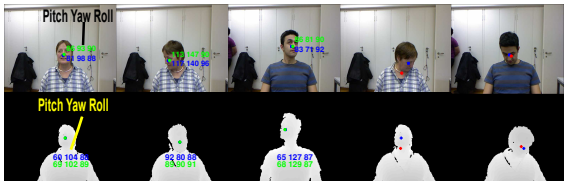
During testing, patches sampled from the test image traverse the trees and cast votes to the Hough space $\mathcal{H}$ based on the location and pose distributions stored in the leaves. The forest-based estimate is then computed by aggregating votes from different patches. The PLEV starts by collecting the votes in our multidimensional Hough space $\mathcal{H}$. We first project all votes on the (x,y) subspace of $\mathcal{H}$, and recover the object center hypothesis $\hat{\mathbf{h}}_\mathbf{d} = (\hat{x}, \hat{y})$ where the maximum is.

We then build a local Hough region $H_r^{\hat{\mathbf{h}}_\mathbf{d}} \subset \mathcal{H}$ for each detection hypothesis $\hat{\mathbf{h}}_\mathbf{d}$. We consider to be in the defined local region only those voting positions which receive at least one vote and are spatially close to the detected maximum. Then, PLEV aggregates all *pose* votes received within $H_r^{\hat{\mathbf{h}}_\mathbf{d}}$, obtaining the distribution of the poses in the Hough region (see Figure 1). Then, a Gaussian KDE is performed on that distribution in order to obtain a smooth probability density function (PDF) for the pose estimation. So, with the PLEV, our HF can cope with the uncertainty of the pose estimation votes.

To further improve the detections, we finally propose to integrate a novel **pose-based backprojection** (BP) strategy to boost the bounding box (BB) estimation using the pose cues. Essentially, we extend the traditional BP strategy [2]. When computing the BP mask, we want to penalize patches that vote not only for different object locations, as in [2], but also for different poses. For more details, see Section 2.3 in the paper.

As a conclusion, we have proposed a new object detection and continuous pose estimation solution using HF. It can successfully detect objects, while the pose is estimated with a probabilistic output using the PLEV. Our method reports state-of-the-art results on 4 different datasets [1, 3, 4, 5]. We show results on cars as well as faces, and using RGB as well as depth images as input. As a HF based approach with simple features, it is efficient. Being a voting-based scheme, it is intrinsically robust to occlusions. While many state-of-the-art approaches need 3D CAD models for the object class of interest during training, our approach is simple in the sense that we are able to learn the model directly from annotated images. Lastly, thanks to our PLEV strategy, we obtain a probabilistic output score, allowing easy integration as a building block in a larger probabilistic framework. Our extension to video-based pose estimation shows how to leverage the temporal continuity in video, even though poses may change from frame to frame. In Figure 2 we show qualitative results for different categories and for different modalities.

(a) Weizmann Cars Viewpoint dataset [3]



(b) Biwi Kinect Head Pose Database [1]

Figure 2: Qualitative results. Ground truth in blue, estimations in green and wrong detections in red.

[1] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *IJCV*, 101(3):437–458, 2013.

[2] J. Gall, N. Razavi, and L. Van Gool. *An Introduction to Random Forests for Multi-class Object Detection*, chapter 11, pages 243–263. Springer, 2012.

[3] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and continuous pose estimation. *IVC*, 30(12):923–933, 2012.

[4] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[5] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. In *ICCV*, 2007.